

2015 NIST Open Machine Translation Challenge

1 INTRODUCTION

The 2015 NIST Open Machine Translation Challenge (OpenMTChallenge'15) continues the ongoing evaluations of human language translation technology. NIST conducts these evaluations in order to support MT research and help advance the state of the art in MT technology. To do this, NIST:

- Defines a set of translation tasks,
- Collaborates with the data providers¹ to provide corpus resources to support research on these tasks,
- Creates and administers formal evaluations of MT technology,
- Provides evaluation utilities to the MT community, and
- Coordinates workshops to discuss MT research findings and results of task performance in the context of these evaluations.

These evaluations provide an important contribution to the direction of research efforts and the calibration of technical capabilities. They are intended to be of interest to all researchers working on the general problem of translating between human languages. To this end, the evaluations are designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible to those wishing to participate.

For the first time, the evaluation is being offered as an on-going challenge rather than a fixed and short duration test to allow flexibility for maximum participation. The 2015 Challenge is planned to be up for six months instead of just one week. A leaderboard tracks the performance. Score feedback is provided for 40% of the test set, and participants can submit up to 5 submissions per day.

The Challenge uses data from the OpenMT15 evaluation, which focuses on informal data genres (SMS/Chat and Conversational Telephone Speech) in Egyptian Arabic and Mandarin Chinese. The task is to translate these data into English.

Participation in the Challenge is open to all researchers who find the task of interest. There is no fee for participation. To participate in the Challenge, sites must officially register with NIST² and agree to the terms specified in the registration form.

2 TRAINING CONDITIONS

MT R&D requires language data resources. System performance and R&D effort are strongly affected by the type and amount of resources used. Therefore, it is important to differentiate training conditions of evaluation as *Constrained Training* and *Unconstrained Training*.

For a submission to be considered in the Constrained Training condition, it must use only specific data listed in the Appendix of this document. Otherwise, it is categorized as Unconstrained.

2.1 CONSTRAINED TRAINING

Systems entered in the Constrained Training condition allow for direct comparisons of different algorithmic approaches. System development must adhere to the following restrictions:

Only data listed in the Appendix may be used for core MT engine development in the Constrained Training condition. **OpenMTChallenge'15 does not place a language specific restriction on the LDC data resources; that is, a site participating in Arabic to English may use Chinese to English data as long as that data is in the list.**

Resources that assist the core engine (such as segmenters, tokenizers, parsers, or taggers) are not subject to the same restriction. If such additional resources are used, they must be listed in the system description.

2.2 UNCONSTRAINED TRAINING

Systems entered in the Unconstrained Training condition when the training data include data that are not in the Constrained Training list. System development must adhere to the following restrictions:

Data must be publicly available, at least in principle.³ This ensures that research results are broadly applicable and accessible to all participants. Participants must specify in their system descriptions what data they used.

3 DATA SETS

3.1 EVALUATION DATA

The OpenMT15 evaluation data will consist of Arabic and Chinese sources. The Arabic data is primarily Egyptian dialect but may have other dialect(s) and/or other language(s). The Chinese data is primarily Mandarin Chinese but may have other dialect(s) and/or other language(s). Each language pair and genre combination has approximately 25K source words. A subset of the 25K (about 5K) reference will have HyTER network annotations.

¹ <http://www ldc.edu> & <http://www sdl.com/research/language-technology>

² <http://openmtchallenge.nist.gov>

³ Data limited to government use, such as the FBIS data, is deemed to be not publicly available and not admissible for system development.

Table 1: Data volume for OpenMT15 test sets.

Language Pair	Genre	Volume (words)
Arabic-to-English	SMS/Chat	~25,000
	CTS	~25,000
Chinese-to-English	SMS/Chat	~25,000
	CTS	~25,000

Each language pair and genre has one gold standard reference. Additionally, HyTER network will be created for approximately 5,000 words for each language pair and genre.

4 INPUT TRACK

Unlike OpenMT15, the OpenMTChallenge'15 offers only for the text input track which consists of the SMS/chat messages as well as the human reference transcripts of the telephone conversations.

The Arabic SMS/chat data may contain a mixture Arabic script, Arabizi, and/or script from other languages, and the human reference transcripts of the Arabic CTS data contain Arabic script and/or script from other languages. The Chinese SMS/chat data may contain a mixture of Chinese characters, pinyin, and/or other languages, and the human reference transcripts of the Chinese CTS data contain Chinese characters and/or script from other languages.

5 PERFORMANCE MEASUREMENT

OpenMTChallenge'15 uses several automatic metrics. BLEU4 is used to determine the site's position in the leaderboard.

- BLEU⁴ – This technique scores a translation according to the N-grams that it shares with one or more reference translations of high quality. In essence, the more co-occurrences, the better the translation. An N-gram, in this context, is simply a *case sensitive* sequence of N tokens. (Words and punctuation are counted as separate tokens.) NIST will compute case-sensitive BLEU scores using NIST's publicly available *mteval* software⁵.
- METEOR⁶ – This technique scores a translation according to word-to-word matches between the system and reference translations but also includes a set of language specific weights tuned to the target language and the ability to incorporate stemming and synonymy.
- TER⁷ – This technique scores a translation according to the number of transformations that must be performed on the system translation such that it has the same word ordering as the reference translation. NIST will compute case-sensitive TER scores using UMD's publicly available *tercom* software⁸. NIST may also compute human-targeted version of TER (HTER).
- HyTER⁹ – This technique is similar to TER but makes use of large networks of reference translations.

6 HONESTY PLEDGE

The OpenMTChallenge'15 is an open evaluation where the participants process the data locally and submit the output to NIST. As such, **the participants have pledged not to probe the evaluation data either by automatic or manual means either via machines or humans to gain additional knowledge about the test data.** Those who are found to violate this pledge will be barred from the Challenge and future NIST evaluations/challenges.

7 CHALLENGE PROCEDURES

The OpenMTChallenge'15 process includes a number of mandatory steps; please see the schedule in Section 12 for the dates for each of these:

- 1 Sign up for an evaluation account. See Section 9.
- 2 Sign the honesty pledge.

⁴ Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu (2001). "Bleu: a Method for Automatic Evaluation of Machine Translation". This report may be downloaded from URL <http://domino.watson.ibm.com/library/CyberDig.nsf/home> (keyword RC22176).

⁵ <ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a-20091001.tar.gz>

⁶ Michael Denkowski and Alon Lavie, "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems", Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation, 2011.

⁷ Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas, 2006.

⁸ <http://www.cs.umd.edu/~snover/tercom/>

⁹ Dreyer, Markus, and Daniel Marcu. "Hyter: Meaning-equivalent semantics for translation evaluation." Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2012.

- 3 Sign LDC's data license agreement and return it to LDC. **Even if not selecting any training data, participants must sign the agreement to receive the evaluation data, which are listed on the agreement.**
- 4 Download the evaluation source data from NIST.
- 5 Perform the evaluation translation. Each site must run its translation system(s) on the entire set.
- 6 Upload the evaluation translations to NIST according to instructions given in Section 9.
- 7 View the submission status. The status of the submission will be posted in the participant's evaluation account.
- 8 Submit a system description (see Section 10).

8 NIST OPENMT DATA FORMAT

Translation systems must be able to process the input source files and produce the translations that meet the OpenMT data format. There are two input tracks in OpenMT15 (see Section 4). The input and output data formats are described below.

8.1 SOURCE FILE FORMAT

The format for the text input track will follow a format similar to one used in previous years with one main difference: *this year each document resides in its own file while in previous years all the documents reside in one file.*

NIST has defined a set of XML tags that are used to format MT source, reference, and translation files for evaluation. All NIST OpenMT source, reference, and translation files have an *xml* extension; their format is defined by the current XML DTD.¹⁰ NIST requires that all submitted translation files are well-formed and valid against the above-mentioned DTD.

A source file contains one single `srcset` element, immediately beneath the root `mteval` element. The `srcset` element has the following attributes:

- `setid`: The dataset.
- `srclang`: The source language. One of: Arabic, Chinese.

The `srcset` element contains one `doc` elements, which have the following attributes:

- `docid`: The document name.
- `genre`: The data genre. One of: `sms`, `chat`, `ctstext`, `ctsaudio`.

Each `doc` element contains several segments (`seg` elements). Each `segment` has a single attribute, `id`, which must be enclosed using double quotes.

One or more segments may be encapsulated inside additional elements, such as (but not limited to) `hl`, `p`, or `poster`. Only the native language text that is surrounded by a `seg` start-tag and its corresponding end-tag is to be translated.

OpenMT15 sample source file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.7.dtd">
<mteval>
  <srcset setid="sample_document_1" srclang="Arabic">
    <doc docid="sample_document_1" genre="sms">
      <seg id="1">ARABIC SENTENCE #1</seg>
      <seg id="2">ARABIC SENTENCE #2</seg>
      ...
    </doc>
  </srcset>
</mteval>
```

The source files will be named as `<base>.arz.su.xml` and `<base>.cmn.su.xml` for Arabic and Chinese source files, respectively.

8.2 REFERENCE FILE FORMAT

A reference file contains one or more `refset` elements, immediately beneath the root `mteval` element. Each `refset` element contains the following attributes:

- `setid`: The dataset.
- `srclang`: The source language. One of: Arabic, Chinese.
- `trglang`: The target language, English.
- `refid`: The current reference.

¹⁰ <ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.7.dtd>

Each `refset` element contains one document, which, in turn, contains the segments. The document elements and their subsequent child elements is exactly the same as described in Section 8.1 above for the source file.

OpenMT15 sample reference file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.7.dtd">
<mteval>
  <refset setid=" sample_document_1" srclang="Arabic" trglang="English" refid="reference01">
    <doc docid="sample_document_1" genre="sms">
      <seg id="1">ENGLISH REFERENCE TRANSLATION #1</seg>
      <seg id="2">ENGLISH REFERENCE TRANSLATION #2</seg>
      ...
    </doc>
  </refset>
</mteval>
```

The reference files will be named as `<base>.eng.su.xml`.

8.3 TRANSLATION (TEST) FILE FORMAT

A translation file contains one `tstset` elements, immediately beneath the root `mteval` element. Each `tstset` element contains the following attributes:

- `setid`: The dataset.
- `srclang`: The source language. One of: Arabic, Chinese.
- `trglang`: The target language, English.
- `sysid`: A name identifying site and system (see Section 9.3.1 for requirements).

The content of each `tstset` element is exactly the same as described previously for the source file format and the reference file format.

OpenMT15 sample translation (test) file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.7.dtd">
<mteval>
  <tstset setid=" sample_document_1" srclang="Arabic" trglang="English" sysid=" NIST_ara2eng_primary_cn">
    <doc docid="sample_document_1" genre="sms">
      <seg id="1">ENGLISH SYSTEM TRANSLATION #1</seg>
      <seg id="2">ENGLISH SYSTEM TRANSLATION #2</seg>
      ...
    </doc>
  </tstset>
</mteval>
```

The translation files will be named as `<base>.eng.su.xml`. Note that the translation files must preserve the value of the id attributes (`setid`, `docid`, `id`) of the corresponding source files so that the translations can be referenced back to the source and reference translations.

9 EVALUATION INSTRUCTIONS

9.1 EVALUATION ACCOUNT

All evaluation activities will be stored in an evaluation account for each participant. Therefore, participants must sign up for an evaluation account. To sign up for an account, go to <https://openmtchallenge.nist.gov>.

After successfully confirming the account, participants will be able to perform various tasks such as

- download the source data using the “Get the data” link at the top of the page
- upload your submission using the “Upload submission” link at the top of the page
- check submission status using the “Status” link also at the top of the page

9.2 SOURCE DATA DOWNLOAD

The evaluation source data will be available as download via the evaluation account. **Inspection and manipulation of the evaluation data via manual or automatic means are prohibited.**

9.3 SUBMISSION REQUIREMENTS

Participants may submit up to 5 valid submissions per day. Submissions that do not pass validation do not count toward the daily submission limit. Each submission is scored upon passing validation. While all documents in each submission are scored, only 40% of the

scores are shown. An average score is computed for each submission, and that score is used to determine a site's best score to date and position in the leaderboard. Each submission must follow the format described below.

9.3.1 File Naming

The system output files must comply with the following naming convention:

`<base>.eng.su.xml`

Where:

- `base`: The base filename of the test file and should match the base of the input test file.

9.3.2 Submission File

All system output files must be tarred into a submission file. The submission file has the following format:

`<testsetID>_<condition[-label]>.tgz`

Where:

- `testsetID`: The ID of the testset used. This is also the name of the data you downloaded.
 - One of: `openmt15-ara2eng-text` or `openmt15-chi2eng-text` (Note: the hyphen in the testset ID and do not confuse with the underscore that is used to separate the fields in the submission filename)
- `condition`: The training condition, referring to Constrained or Unconstrained training
 - One of: `cn`, `un`
- `label`: The label is optional and is defined by the participant (e.g. to help the participant to remember the submission or whatever the participant wants). Note: the hyphen and not the underscore separating `condition` and `label` if there is a `label`.

9.3.3 Submission Instructions

The submission file must be named as described in and packaged as follows:

- Place the system output files in a clean directory
 - `cp <system output xml files> <some clean directory>`
- `cd` into the directory, `tar` and `zip` system files with the name as described in 9.3.2
 - `cd <some clean directory>`
 - `tar zcfv openmt15-ara2eng-text_un-1.tgz *xml`
- Upload your submission using your evaluation account

10 SYSTEM DESCRIPTIONS

Participants are required to submit system descriptions of the MT systems used for their submissions. Please use NIST's template¹¹ for system descriptions. System descriptions should be submitted in text format, and the file name should reflect the site ID. System description is due one month after the initial system submission or by May 18, 2016 whichever is first. Participants are free to submit as many revisions as they wish. If no system description is received for a given site at the conclusion of the Challenge, that site's results will be removed from the leaderboard and all NIST report(s).

11 GUIDELINES FOR PUBLICATION OF RESULTS

NIST Multimodal Information Group's MT evaluations follow an open model to promote interchange with the outside world. At the conclusion of the evaluation cycle, NIST will create a report that documents the evaluation. The report will be posted on the NIST web space and will identify the participants and the scores from various metrics achieved for each language pair and training condition. Results from the human assessments may also be posted.

The report that NIST creates should not be construed or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

11.1 RULES GOVERNING PUBLICATION OF EVALUATION RESULTS

The rules governing the publication of NIST OpenMTChallenge'15 results are similar to those used in OpenMT's.

- Participants must refrain from publishing results and/or releasing statements of performance until the official OpenMTChallenge'15 results are posted by NIST. Statements of performance may not claim winning or be perceived as a ranking amongst other participants.

¹¹ http://www.nist.gov/itl/iad/mig/upload/OpenMTChallenge15_SysDescTemplate.txt

- Participants are free to publish results for their own system, but participants must not publicly compare their results with other participants (ranking, score differences, etc.) without explicit written consent from the other participants. Publications should point to the NIST report as a reference.¹²
- NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST, or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.
- All publications must contain the following NIST disclaimer:

NIST serves to coordinate the NIST OpenMT evaluations in order to support machine translation research and to help advance the state-of-the-art in machine translation technologies. NIST OpenMT evaluations are not viewed as a competition, as such results reported by NIST are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U.S. Government.

- Linguistic resources used in building systems for OpenMTChallenge'15 should be referenced in the system description. Corpora should be given a formal citation, like any other information source. LDC corpus references should adopt the following citation format:

Author(s), Year. Catalog Title (Catalog Number). Linguistic Data Consortium, Philadelphia PA.

For example:

Xiaoyi Ma et al, 2005. Arabic News Translation Text Part 1 (LDC2004T17). Linguistic Data Consortium, Philadelphia PA.

12 SCHEDULE (TENTATIVE)

- Challenge opens: November 18, 2015
- System description: one month after first submission or May 18, 2016 whichever is first
- Challenge closes: May 30, 2016

¹² This restriction exists to ensure that readers concerned with a particular system's performance will see the entire set of participants and tasks attempted by all researchers.